DMP du projet "Unraveling breast cancer stem cell enhancer landscape remodeling in response to BET protein inhibitors: characterization of new therapeutic targets for triple-negative breast cancer"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "Science Europe : modèle structuré" fourni par Science Europe.

Renseignements sur le plan

Titre du plan

DMP du projet "Unraveling breast cancer stem cell enhancer landscape remodeling in response to BET protein inhibitors: characterization of new therapeutic targets for triple-negative breast cancer"

Version

Version initiale

Objet/périmètre du plan

Research products

1. Epigenomic profiles of bCSCs (and non-bCSC lineages) in response to BETi treatments:
    a. RNA-seq;
    b. CUT&RUN for histone modifications and BRD proteins;
    c. Capture Hi-C for promoter-mediated chromatin interactions.
2. Single-cell RNA and ATAC-seq profiles of bCSC differentiation and response to repression of candidate factors.
3. *In vivo* validation of potential targets uncovered to be responsible for bCSC identity.

1. Data description and collection or re-use of existing data

Epigenomic profiles of bCSCs (and non-bCSC lineages) in response to BETi treatments.

1a. How will new data be collected or produced and/or how will existing data be re-used?

First, to validate our approach with the limited existing data, we will use SUM159 cells. To obtain more physiopathologically relevant information, we will use in parallel patient-derived xenografts from two TNBC patient samples issued from a PDX bank generated within the consortium. Specifically, we will use models (CRCM404 and CRCM434) whose genome and transcriptome have been extensively characterized, and for which a bCSC population was readily isolated, whose differentiation was shown to be induced by JQ1 treatment. For these three sources (SUM159, CRCM404, CRCM434), bCSCs and non-bCSCs will be sorted by FACS for functionally validated bCSC cell-surface markers (CD44+/CD24-) and high ALDH1 activity, forming six different cell preparations, which will be used in biological replicates for the following low cell number epigenomic assays:

- PolyA-RNA-seq to characterize steady-state gene transcription;
- Total (rRNA-depleted) RNA-seq to identify eRNAs and obtain a better view of any differences in nascent transcription;
- ATAC-seq to determine the sites of accessible chromatin;
- CUT&RUN for H3K27ac and H3K4me1, the histone

modifications most commonly associated with enhancers (and active promoters, to some extent);

- Promoter Capture-LoC to identify the target genes of distal regulatory elements.

1b. What data (for example the kind, formats, and volumes) will be collected or produced?

Raw Illumina files will be processed for quality and converted to standard fastq files, comprising the “raw” sequencing data for archiving. RNA-seq will be processed by standard tools (Star, HT-Seq) to quantify steady-state expression of each gene in each condition (an RPKM value assigned to each annotated gene). Differential analysis will be performed by DESeq2. CUT&RUN data will be processed with minor modifications to standard ChIP-seq pipelines (e.g. to trim specialized adapters) (MACS2 and IDR calling on duplicate experiments) to identify high-confidence discrete peaks. Different combinations of intersections of these called peaks at intergenic and intronic regions will be used to call putative enhancers of differing confidence, and DESeq2 will be used to assess quantitative as well as qualitative differences between bCSCs and non-bCSCs. Motif searches with tools such as HOMER will be used to look for key transcription factors more specific to bCSCs. Promoter-mediated interactions will be called from the Capture-LoC datasets using the ChiCMaxima tool developed in the consortium.

Single-cell RNA and ATAC-seq profiles of bCSC differentiation and response to repression of candidate factors.

1a. How will new data be collected or produced and/or how will existing data be re-used?

To better understand the bCSC differentiation process, we will purify bCSCs as before, then place the cells for 24 hours in culture. Cells will then be isolated and subjected to simultaneous single-cell RNA-seq and ATAC-seq with the 10X Genomics Multiome Chromium system. We will implement the newly developed ECCITE-seq (expanded CRISPR-compatible cellular indexing of transcriptomes and epitopes) method to simultaneously assess the role of perturbing candidate genes on bCSC maintenance and transcriptional output. We will transduce SUM159 cells to generate two derivative lines expressing orthologous fluorescent reporters (i.e. BFP and RFP) fused to Cas9. These will be co-cultured as a population of 5% ALDH+/CD44+/CD24- BFP+ bCSCs and 95% ALDH-/CD44+/CD24+ RFP+ non-bCSCs, then transduced with a library of gRNAs targeting ~100 candidate genes. “Newly” differentiated (ALDH-/CD44+/CD24+ BFP+) and “stably” differentiated (ALDH-/CD44+/CD24+ RFP+) cells will be isolated by FACS and single cells will be subjected to an adapted scRNA-seq which simultaneously captures the gRNA species present within each cell.

1b. What data (for example the kind, formats, and volumes) will be collected or produced?

Using the steady-state bulk transcriptome and chromatin accessibility profiles of bCSCs and non-bCSCs generated in other work packages as “end-point” references, the scRNA-seq and scATAC-seq data will be clustered with UMAP, and trajectories will be added with pseudotime and RNA velocity analyses.

***In vivo* validation of potential targets uncovered to be responsible for bCSC identity.**

1a. How will new data be collected or produced and/or how will existing data be re-used?

We will knock down a small number (~3-5) of the best candidate genes using CRISPRi, using lentiviral constructs developed within the consortium. We will perform limited dilution analysis to evaluate the frequency of bCSCs in each PDX model after candidate gene knockdown. Briefly, we will perform an extreme limiting dilution assay (ELDA) by xenografting in mice a serial dilution of mCherry+/BFP+ cells (usually 50K, 5K, or 500 cells).

1b. What data (for example the kind, formats, and volumes) will be collected or produced?

Cell counts and cytometric analysis over a time course, generating simple text/Excel files.

2. Documentation and data quality

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

All protocols for the experiments resulting in data generation and subsequent analysis are stored as Word files. The direct metadata included for the sequencing outputs are sequence quality scores (part of the fastq format) and the identifier for the experiment type and condition, as well as unique cell IDs for the single-cell sequencing data. All other metadata will be included in keeping with the Gene Expression Omnibus guidelines.

2b. What quality control measures will be used?

All sequencing data are checked for quality in terms of sequence complexity and read calling quality. Biological duplicates are cross-compared for reproducibility, and used for subsequent statistical analyses.

3. Storage and backup during the research process

3a. How will data and metadata be stored and backed up during research?

All raw files (namely the fastq files for sequencing data will be archived and backed up on the secure servers within the IGBMC and CRCM. Both institutes have the appropriate infrastructure for secure storage of petabytes of data.

3b. How will data security and protection of sensitive data be taken care of during the research

The secure servers are only accessible from the intranet, or via VPN connection, all of which is password-protected. All data are systematically backed up daily.

4. Legal and ethical requirements, code of conduct

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

Not applicable.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

As academic research, all data will be made publicly available on publication. In the case of potential financial exploitation of the project results, the project coordinators will keep ties with the SATT Connectus, who will advise on the procedure for patenting and IP protection.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

Public dissemination of all methods, protocols and raw data will ensure that our results can be assessed for reproducibility and lack of data tampering.

5. Data sharing and long-term preservation

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Data will be made available to the scientific community on publication of the results of the different studies. No restrictions to data sharing are foreseen at the moment.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Sequencing data will be deposited on the Genome Expression Omnibus, and all code will be made available on github. Processed data results will be stored long-term on the IGBMC server, and will be available on request, if not already present as supplemental tables within publications.

5c. What methods or software tools are needed to access and use data?

The same software tools required to process the data in the first place, all of which are publicly available.

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

All data published on external public repositories will automatically receive a digital identifier.

6. Data management responsibilities and resources

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

Ultimate responsibility will lie with the project coordinators, who will ensure that the appropriate protocols about data generation and backup are being followed. Ensuring the smooth running of the server infrastructure, including security and backup, are already within the remit of the permanent IGBMC IT staff, headed by Julien Seiler.

6b. What resources (for example financial and time) will be dedicated to

	data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)? No extra resources are required, since all used software is free and publicly available, and the IGBMC server maintenance is assured by permanent staff members.
Domaines de recherche (selon classification de l'OCDE)	Biological sciences (Natural sciences)
Langue	eng
Date de création	2023-11-06
Date de dernière modification	2023-11-06
Identifiant	DMP:PLBIO22-118
Type d'identifiant	DOI
Licence	Creative Commons Attribution Non Commercial 4.0 International

Renseignements sur le projet

Titre du projet	Unraveling breast cancer stem cell enhancer landscape remodeling in response to BET protein inhibitors: characterization of new therapeutic targets for triple-negative breast cancer

Résumé

Our understanding of breast cancer is being transformed by exploring clonal diversity, cell plasticity, drug resistance, and causation within an evolutionary framework. Tumor cell heterogeneity relies on cell-state transition dynamics, which largely impact therapeutic resistance. A major part of this plasticity appears to be conferred by cancer stem cells (CSCs), a cell subpopulation of solid tumors that self-renew and can differentiate to various cell types comprising the bulk of the cancer. Drugs targeting CSCs are a promising avenue to eliminate drug-tolerant state cells and avoid tumor relapse. Very recently, we discovered that pan-inhibition of the BET (Bromodomain and Extra-Terminal motif) proteins promoted differentiation of breast CSCs (bCSCs) and reduced tumorigenicity both in vitro and in vivo. Despite mounting evidence that such treatment perturbs enhancers responsible for tumor progression, it is unclear which enhancers and gene targets are the most important targets, nor in which cellular contexts they are the most critical, especially since nearly all enhancer characterization has been performed in bulk cells. Understanding the specificity of CSC vs non-CSC gene-enhancer networks will thus be important in understanding their interplay within a heterogeneous tumor environment and their response to epigenetic drugs, as well as informing more targeted therapeutic strategies. Our consortium proposes to combine expertise in epigenomics, oncology, stem cell biology and chemical biology to address the following questions, always with a view to better define triple-negative breast cancer (TNBC) therapies:

- Which enhancers and genes determine bCSC and non-bCSC identities in vivo?
- Which epigenetic changes accompany bCSC differentiation?
- What are the respective contributions of the BET members and of each their tandem bromodomains to bCSC identity?
- Which molecular pathways underlie the responses of TNBC to (selective) BET inhibitors?

Sources de financement

- Institut National Du Cancer : PLBIO22-118

Date de début 2022-10-20

Date de fin 2026-10-19

Produits de recherche :

1. Epigenomic profiles of bCSCs (and non-bCSC lineages) in response to BETi treatments

Contributeurs

Nom	Affiliation	Rôles
Sexton Thomas		• Coordinateur de projet • Personne contact pour les données • Responsable de la production ou de la collecte des données • Responsable du plan

DMP du projet "Unraveling breast cancer stem cell enhancer landscape remodeling in response to BET protein inhibitors: characterization of new therapeutic targets for triple-negative breast cancer"

1. Description des données et collecte ou réutilisation de données existantes

1.1 Description générale du produit de recherche

Nom	Epigenomic profiles of bCSCs (and non-bCSC lineages) in response to BETi treatments
Description	*1a. How will new data be collected or produced and/or how will existing data be re-used?* First, to validate our approach with the limited existing data, we will use SUM159 cells. To obtain more physiopathologically relevant information, we will use in parallel patient-derived xenografts from two TNBC patient samples issued from a PDX bank generated within the consortium. Specifically, we will use models (CRCM404 and CRCM434) whose genome and transcriptome have been extensively characterized, and for which a bCSC population was readily isolated, whose differentiation was shown to be induced by JQ1 treatment. For these three sources (SUM159, CRCM404, CRCM434), bCSCs and non-bCSCs will be sorted by FACS for functionally validated bCSC cell-surface markers (CD44+/CD24-) and high ALDH1 activity, forming six different cell preparations, which will be used in biological replicates for the following low cell number epigenomic assays: • PolyA-RNA-seq to characterize steady-state gene transcription; • Total (rRNA-depleted) RNA-seq to identify eRNAs and obtain a better view of any differences in nascent transcription; • ATAC-seq to determine the sites of accessible chromatin; • CUT&RUN for H3K27ac and H3K4me1, the histone modifications most commonly associated with enhancers (and active promoters, to some extent); • Promoter Capture-LoC to identify the target genes of distal regulatory elements. *1b. What data (for example the kind, formats, and volumes) will be collected or produced?* Raw Illumina files will be processed for quality and converted to standard fastq files, comprising the “raw” sequencing data for archiving. RNA-seq will be processed by standard tools (Star, HT-Seq) to quantify steady-state expression of each gene in each condition (an RPKM value assigned to each annotated gene). Differential analysis will be performed by DESeq2. CUT&RUN data will be processed with minor modifications to standard ChIP-seq pipelines (e.g. to trim specialized adapters) (MACS2 and IDR calling on duplicate experiments) to identify high-confidence discrete peaks. Different combinations of intersections of these called peaks at intergenic and intronic regions will be used to call putative enhancers of differing confidence, and DESeq2 will be used to assess quantitative as well as qualitative differences between bCSCs and non-bCSCs. Motif searches with tools such as HOMER will be used to look for key transcription factors more specific to bCSCs. Promoter-mediated interactions will be called from the Capture-LoC datasets using the ChiCMaxima tool developed in the consortium.
Mots clés (texte libre)	
Identifiant pérenne	GEO accession
Type d'identifiant	DOI
Contient des données personnelles ?	Non
Contient des données sensibles ?	Non
Prend en compte des aspects éthiques ?	Oui

1.2 Est-ce que des données existantes seront réutilisées ?

Justification	Non

1.3 Comment seront produites/collectées les nouvelles données ?

Nom de la méthode	Sequencing-based assays
Description	First, to validate our approach with the limited existing data, we will use SUM159 cells. To obtain more physiopathologically relevant information, we will use in parallel patient-derived xenografts from two TNBC patient samples issued from a PDX bank generated within the consortium. Specifically, we will use models (CRCM404 and CRCM434) whose genome and transcriptome have been extensively characterized, and for which a bCSC population was readily isolated, whose differentiation was shown to be induced by JQ1 treatment. For these three sources (SUM159, CRCM404, CRCM434), bCSCs and non-bCSCs will be sorted by FACS for functionally validated bCSC cell-surface markers (CD44+/CD24-) and high ALDH1 activity, forming six different cell preparations, which will be used in biological replicates for the following low cell number epigenomic assays: • PolyA-RNA-seq to characterize steady-state gene transcription; • Total (rRNA-depleted) RNA-seq to identify eRNAs and obtain a better view of any differences in nascent transcription; • ATAC-seq to determine the sites of accessible chromatin; • CUT&RUN for H3K27ac and H3K4me1, the histone modifications most commonly associated with enhancers (and active promoters, to some extent); • Promoter Capture-LoC to identify the target genes of distal regulatory elements.

2. Documentation et qualité des données

2.1 Quelles métadonnées et quelle documentation (par exemple mode d'organisation des données) accompagneront les données ?

Description	All protocols for the experiments resulting in data generation and subsequent analysis are stored as Word files. The direct metadata included for the sequencing outputs are sequence quality scores (part of the fastq format) and the identifier for the experiment type and condition, as well as unique cell IDs for the single-cell sequencing data. All other metadata will be included in keeping with the Gene Expression Omnibus guidelines.

2.2 Quelles seront les méthodes utilisées pour assurer la qualité scientifique des données ?

3. Exigences légales et éthiques, code de conduite

3.1 Quelles seront les mesures appliquées pour assurer la protection des données à caractère personnel ?

3.2 Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

3.3 Quels sont les aspects éthiques à prendre en compte lors de la collecte des données ?

Description	Public dissemination of all methods, protocols and raw data will ensure that our results can be assessed for reproducibility and lack of data tampering.

4. Traitement et analyse des données

4.1 Comment et avec quels moyens seront traitées les données ?

5. Stockage et sauvegarde des données pendant le processus de recherche

5.1 Comment les données seront-elles stockées et sauvegardées tout au long du projet ?

Besoins de stockage	Sequencing data will be deposited on the Genome Expression Omnibus, and all code will be made available on github. Processed data results will be stored long-term on the IGBMC server, and will be available on request, if not already present as supplemental tables within publications.

6. Partage des données et conservation à long terme

6.1 Comment les données seront-elles partagées ?

Modalités de partage	Sequencing data will be deposited on the Genome Expression Omnibus, and all code will be made available on github. Processed data results will be stored long-term on the IGBMC server, and will be available on request, if not already present as supplemental tables within publications.

6.2 Comment les données seront-elles conservées à long terme ?

Justification	All raw files (namely the fastq files for sequencing data will be archived and backed up on the secure servers within the IGBMC and CRCM. Both institutes have the appropriate infrastructure for secure storage of petabytes of data.